NLP: Relation Classification

- Transfer learning
- Explainability

Alexander Sifel (01427034),
Ana Terović (01427034),
Lukas Hofstetter (51820745)

Datasets and Preprocessing

- FoodDisease
  - 4 features:
    food_entity, disease_entity, sentence, disease_doid
  - 2 categories:
    is_cause, is_treat
  - 609 instances
    is_cause 141T-464F, is_treat 322T-286F

- CrowdTruth Medical Relation Extraction
  - 17 features:
    SID, relation, sentence_relation_score, crowd,
    baseline, expert, test_partition,
    term1, b1, e1, term2, b2, e2, sentence,
    term1_UMLS, term2_UMLS, UMLS_seed_relation
  - 2 categories:
    is_cause, is_treat
  - 3985 instances each class

Datasets and Preprocessing

- FoodDisease
  - 4 features:
    food_entity, **disease_entity**, **sentence**, disease_doid
  - 2 categories:
    is_cause, **is_treat**
  - 609 instances
    is_cause 141T-464F, is_treat 322T-286F

- CrowdTruth Medical Relation Extraction
  - 17 features:
    SID, relation, sentence_relation_score, crowd,
    baseline, expert, test_partition,
    term1, b1, e1, **term2**, b2, e2, **sentence**,
    term1_UMLS, term2_UMLS, UMLS_seed_relation
  - 2 categories:
    is_cause, is_treat
  - 3985 instances each class

Datasets and Preprocessing

- FoodDisease
  - 5 features:
    food_entity, disease_entity, sentence, is_cause, is_treat
  - 609 instances

- CrowdTruth Medical Relation Extraction
  - 5 features:
    term1, term2, sentence, is_cause, is_treat
  - 7670 instances

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  “The disorder can present with a migratory ture of ARTHRITIS with many other features like HEART PROBLEMs, skin rash, gait abnormality and skin nodules.”

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  “the disorder can present with a migratory ture of arthritis with many other features like heart problems, skin rash, gait abnormality and skin nodules.”

lowercase

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  “the disorder can present with a migratory ture of TERMONE with many other features like TERMTWOs , skin rash, gait abnormality and skin nodules.”

entity replacement

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  ['the', 'disorder', 'can', 'present', 'with', 'a', 'migratory', 'ture', 'of', 'TERMTWO', 'with', 'many', 'other', 'features', 'like', 'TERMONE', 'skin', 'rash', 'gait', 'abnormality', 'and', 'skin', 'nodules']

tokenization

nltk.RegexpTokanizer(r‘\w+’)

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

['disorder', 'present', 'migratory', 'ture', 'TERMTWO', 'many', 'features', 'like', 'TERMONE', 'skin', 'rash', 'gait', 'abnormality', 'skin', 'nodules']

stopword removal

nltk.corpus.stopwords.words(‘english’)

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  ['disord' , 'present', 'migratori', 'ture', 'termtwo', 'mani', 'featur' , 'like', 'termon' , 'skin', 'rash', 'gait', 'abnorm' , 'skin', 'nodul']

  stemming

  nltk.PorterStemmer().stem()

  ['disorder', 'present', 'migratory', 'ture', 'termtwo', 'many', 'feature' , 'like', 'termone', 'skin', 'rash', 'gait', 'abnormality', 'skin', 'nodule']

  lemmatization

  nltk.stem.WordNetLemmatizer().lemmatize()

Datasets and Preprocessing

- CrowdTruth Medical Relation Extraction

  [
  'termtwo', 'featur' , 'like', 'termon' ,
  ]

shortest dep path
with
stopword removal

networkx.Graph()
nltk.corpus.stopwords.words('english')

Datasets and Preprocessing

- FoodDisease
  - 9 features:
    food_entity, disease_entity, sentence,
    tokens, tokens_stem, tokens_lemma,
    sdp, sdp_tokens_lemma, sdp_joined
  - 2 classes:
    is_cause, is_treat
  - 588 instances

- CrowdTruth Medical Relation Extraction
  - 7 features:
    term1, term2, sentence,
    tokens, tokens_stem, tokens_lemma,
    sdp_tokens_lemma
  - 2 classes:
    is_cause, is_treat
  - 7670 instances

Relation extraction dataset

- Available
  - FoodDisease
  - CrowdTruth Medical Relation Extraction

- Difficulty
  - CrowdTruth dataset has very bad quality
  - Almost all **is_treat** instances are wrongly labeled
  - Linear models unable to converge

- Approach
  - Use only FoodDisease
  - Replace entities with placeholders (term 1: **influence**, term2: **condition**)
  - Calculate Shortest Dependency Path (SDP) for comparison
  - In total 588 rows after preprocessing (132 **is_cause**, 313 **is_treat**)
  - Keep same 10% of samples for testing to make results comparable

Relation (multi-label) classification

- Baseline
  - BoW Naive Bayes classifier

BoW + NB	Precision	Recall	F1	Support
is_cause	1.00	0.07	0.13	14
is_treat	0.81	0.88	0.84	33
micro_avg	0.81	0.64	0.71	47
macro_avg	0.90	0.48	0.49	47

- Improvement ideas
  - BERT features + traditional model
  - Finetuned classifier with BERT encoder

- Choosing BERT
  - Training data should have similar domain as ours: Medicine, Biology
  - We use the **emilyalsentzer/Bio_ClinicalBERT** checkpoint
  - https://arxiv.org/abs/1904.03323

BERT Features + Linear SVC

Features from sentences

- Create padded/truncated token IDs from sentence with tokenizer and encode with **BERT**
- Get last hidden state of the **CLS** special token embedding (first embedding of output sequence)
- For SDP and full sentence

Train model

- Linear SVC
- C: [0.01, 0.1, 1]
- 10-fold cross validation

BERT Features + Linear SVC - results

- Full sentence input better than SDP (same as with BoW + NB)
- Balances precision/recall for **is_cause** (BoW + NB has no recall)
- Accuracy similar to BoW + NB
- Bit higher F1 but much less precision

BERT + SVM	Precision	Recall	F1
is_cause	0.57	0.86	0.69
is_treat	0.86	0.76	0.81
micro_avg	0.74	0.79	0.76
macro_avg	0.72	0.81	0.75

BERT + SVM (SDP)	Precision	Recall	F1
is_cause	0.47	0.50	0.48
is_treat	0.76	0.67	0.71
micro_avg	0.66	0.62	0.64
macro_avg	0.61	0.58	0.60

Finetuned BERT classifier

- Use full sentences (want to learn full context)
- BERT base **CLS** embedding into linear layer with 768 inputs and 2 outputs
- Unlock BERT layer gradients for finetuning
- Multi-label classification -> binary cross-entropy loss
- Early stopping with patience monitoring validation loss
- 10% of train set for cross validation

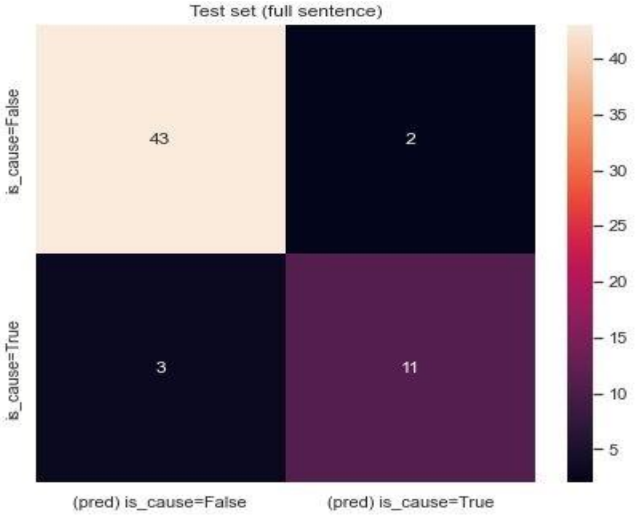
Finetuned BERT classifier - results

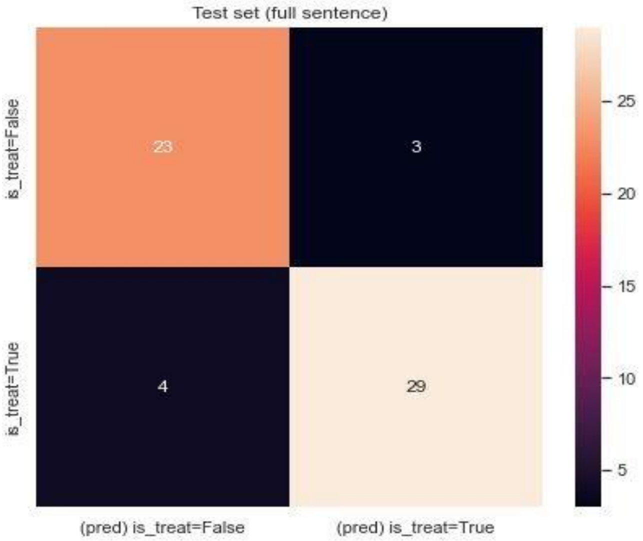
- Finetuning delivers best results by far
- Some (tricky) test examples predictions are still predicted wrongly
- Precision is slightly better than baseline, with much improved recall
- Very high F1 -> try to trade for extra precision by increasing classification threshold

Finetuned	Precision	Recall	F1
is_cause	0.85	0.79	0.81
is_treat	0.91	0.88	0.89
micro_avg	0.89	0.85	0.87
macro_avg	0.88	0.83	0.85

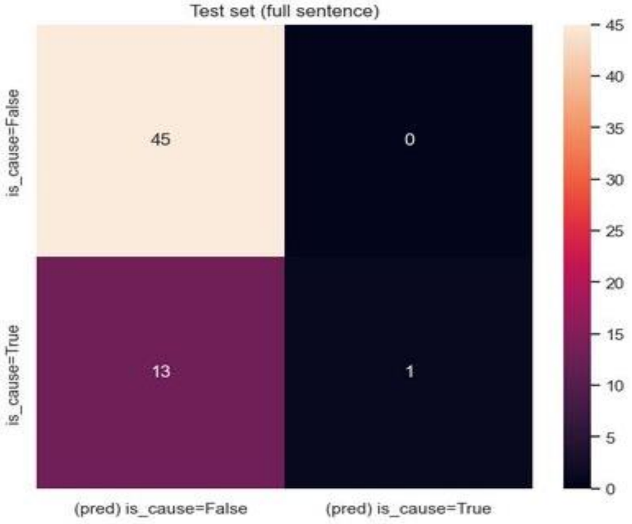
Finetuned BERT Classifier - results

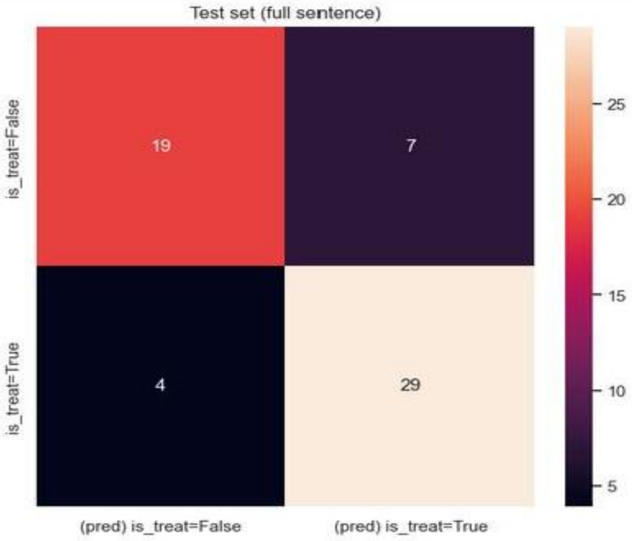
Finetuned





BoW + NB





Finetuned BERT Classifier - results

is_cause false positives:

1. *since <u>influence</u> has been related to the development of chronic <u>condition</u> prevalent in the western world, the use of sweeteners has gradually increased worldwide over the last few years.*

   -> mislabeled, predicted rightly

2. *<u>condition</u> (brd) is a major cause of morbidity and mortality in <u>influence</u> cattle.*

   -> has all the parts, but the relation direction is not right

is_treat false positives:

1. *however, the validity of <u>influence</u> as a treatment for <u>condition</u> (ra), an autoimmune disorder, has not been confirmed yet*

   -> confusion by counterfactual phrasing

2. *abundant studies have highlighted the protective effects of docosahexaenoic acid (dha), in the form of glycerolipids (glycerophosphatides and triglycerides) and dha-ethyl esters (dha-ee) in <u>condition</u> (ad); however, <u>influence</u> (epa) has rarely been implicated*

   -> confusion by counterfactual phrasing

3. *while <u>influence</u> have been shown to exhibit serious side effects, and bioactive compounds from plant-based functional foods have been demonstrated to be active in the treatment of <u>condition</u> with only minimal side effects.*

   -> relation is in there, but not really related to influence

Summary: Classification

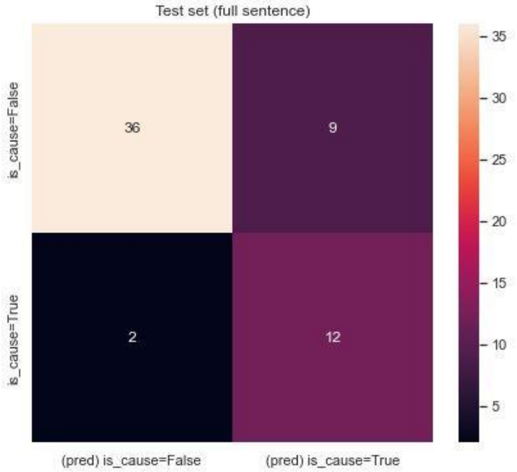
- BERT features make an interesting replacement for BoW
- Finetuning gives generally the best performance
- High precision is important, do not want to pollute knowledge base with false positives -> baseline still very competitive

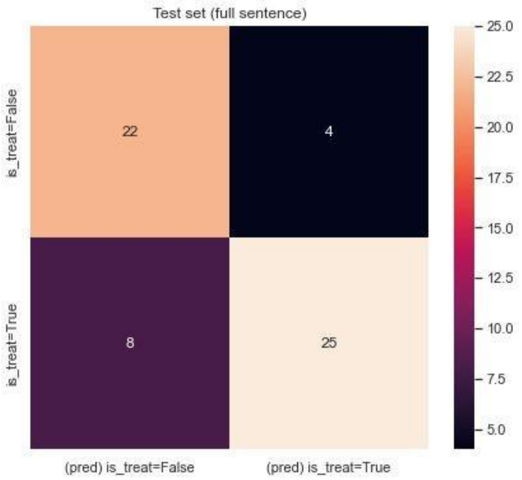
Precision	BoW + NB	BERT + SVM (SDP)	BERT + SVM	Finetuned
is_cause	1.00	0.47	0.57	0.85
is_treat	0.81	0.76	0.86	0.91
micro_avg	0.81	0.66	0.74	0.89
macro_avg	0.90	0.61	0.72	0.88

F1	BoW + NB	BERT + SVM (SDP)	BERT + SVM	Finetuned
is_cause	0.12	0.48	0.69	0.81
is_treat	0.84	0.71	0.81	0.89
micro_avg	0.71	0.64	0.76	0.87
macro_avg	0.49	0.60	0.75	0.85

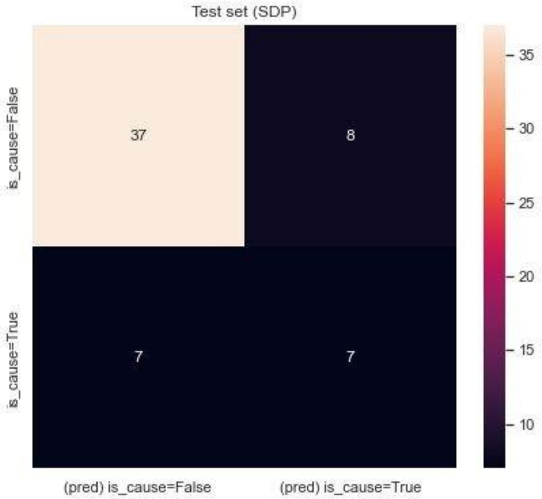
BERT Features + Linear SVC

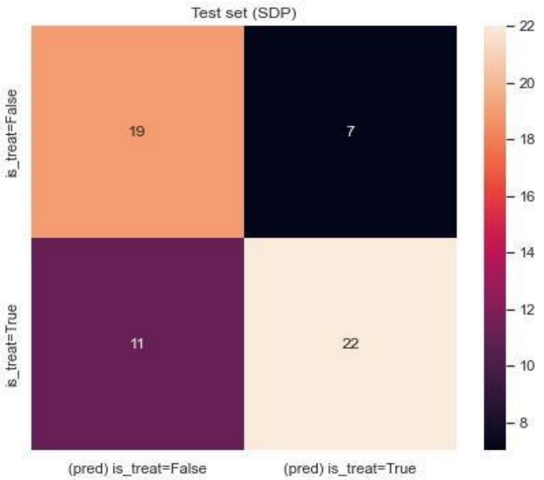
BERT (full)



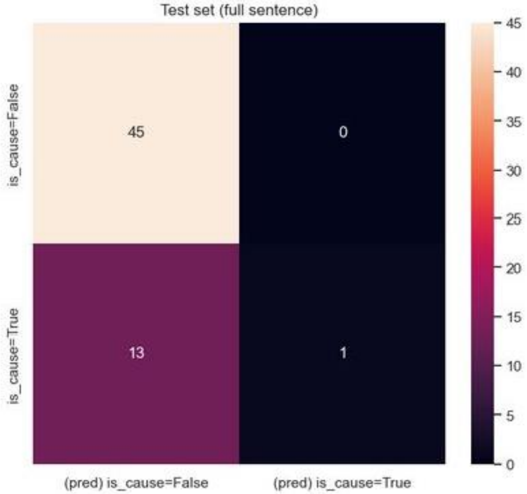


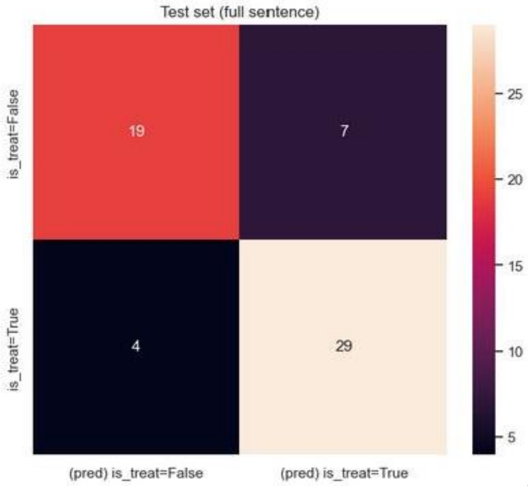
BERT (SDP)





Baseline (full)





Main Findings & Conclusions

Transfer knowledge

Transfer learning from a BERT model works even with a small dataset

Garbage in - garbage out

If the dataset is bad enough your model might not converge at all

POTATO
explainable model

- **rule based system + ML**
  ML is used to learn and generate the rules
- human-in-the-loop learning, **HITL of rules**
- **idea**:
  - subgraphs as features
  - generate subgraphs onlt up to a certain edge number; min_edge, max_edge
  - suggest rules based on feature importance

is_treat

Trainer 1

- **min_edge = 0**
  -> we can have tokens as rules
- top 10 features:

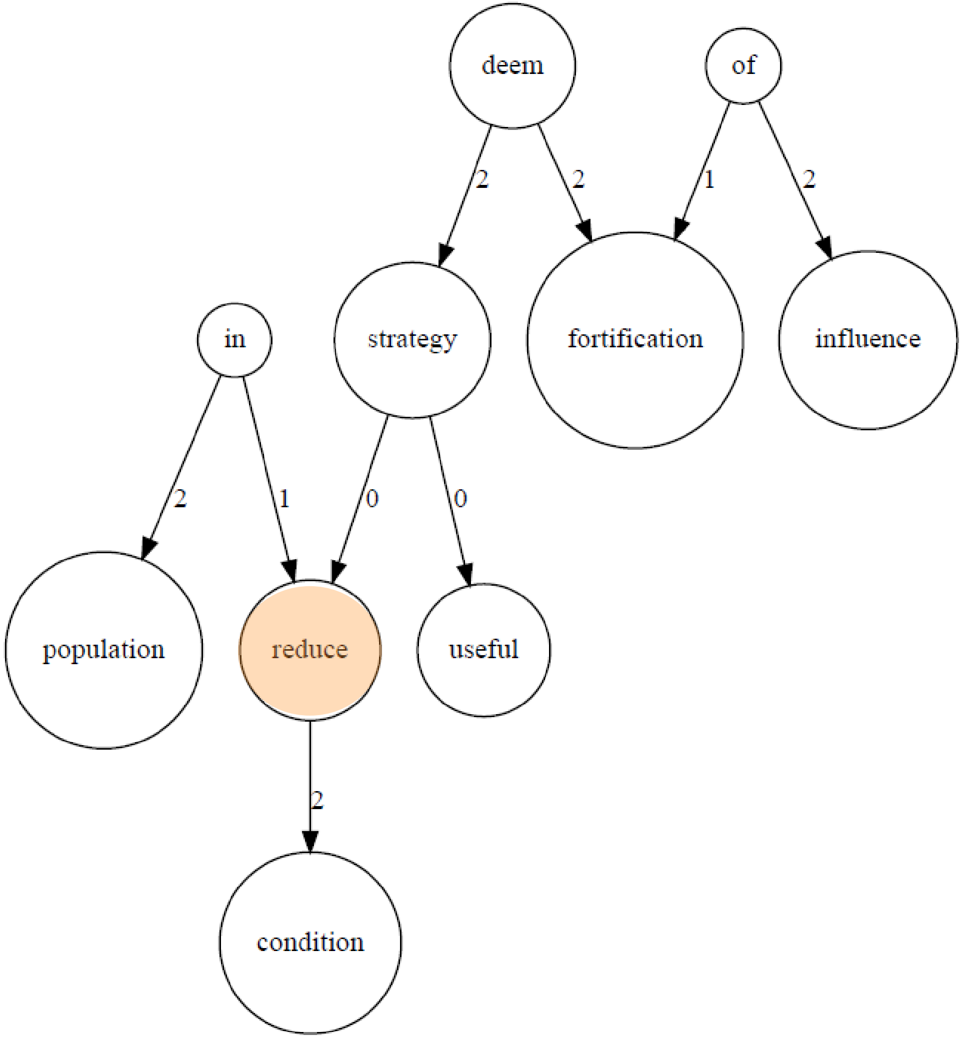
	Feature	Precision	Recall	Fscore
0	[(u_114 / reduce)]	0.866667	0.185714	0.305882
1	[(u_227 / decrease)]	0.833333	0.035714	0.068493
2	[(u_76 / against)]	0.939394	0.110714	0.198083
3	[(u_108 / prevent)]	0.931034	0.096429	0.174757
4	[(u_57 / improve)]	0.950000	0.067857	0.126667
5	[(u_117 / component)]	0.950000	0.067857	0.126667
6	[(u_138 / compound)]	0.913043	0.075000	0.138614
7	[(u_486 / low)]	0.733333	0.078571	0.141935
8	[(u_421 / treat)]	0.904762	0.067857	0.126246
9	[(u_76 / against :2 (u_21 / condition))]	1.000000	0.064286	0.120805

is_treat

reduce

- appears in 60 sentences
- example:

„fortification of influence is
 deemed a useful strategy to reduce
condition in populations”

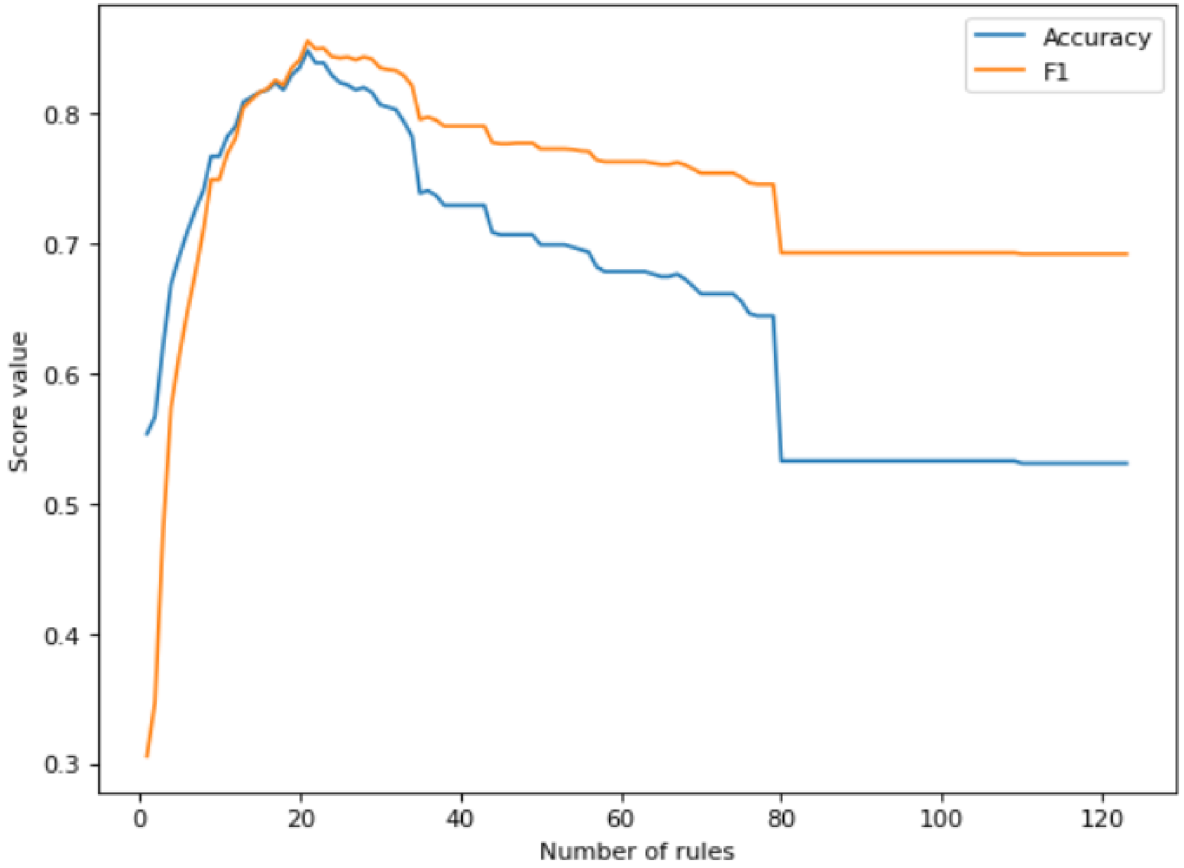


is_treat

Model

- trainer gave 118 features
- ruleset: increment by 1 rules ordered by feature importance
- best results: 20 features

is_treat	POTATO min_edge = 0, 4lang
accuracy	0.77
precision	0.8571
recall	0.7272
F1	0.7869



is_treat

Trainer 2

- **min_edge = 1**
  -> we don't want tokens
- top 10 features:

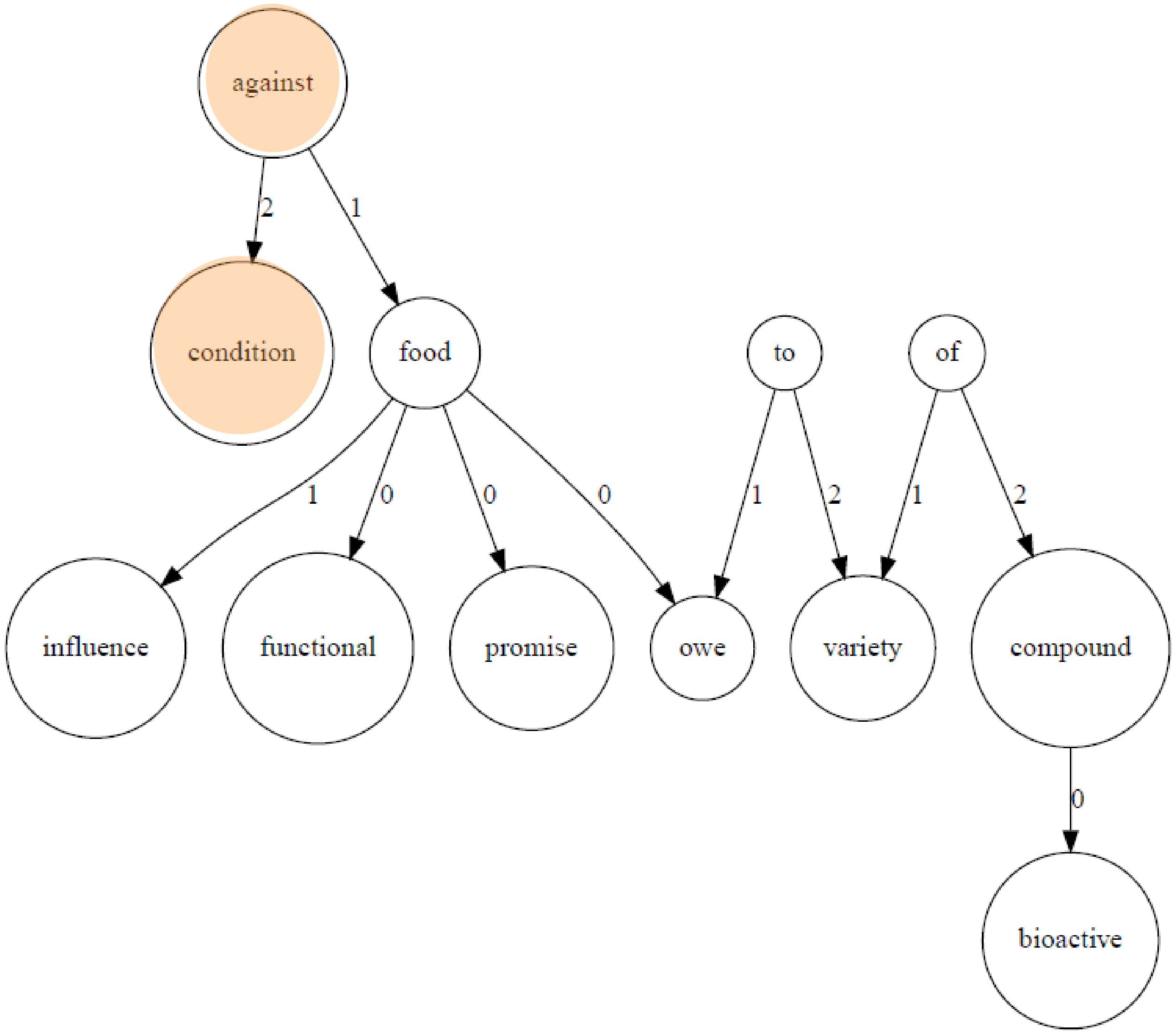
	Feature	Precision	Recall	Fscore
0	[(u_76 / against :2 (u_21 / condition))]	1.000000	0.064286	0.120805
1	[(u_4 / with :2 (u_8 / COORD) :1 (u_89 / ass...	0.692308	0.032143	0.061433
2	[(u_12 / of :1 (u_681 / prevention))]	0.916667	0.039286	0.075342
3	[(u_12 / of :1 (u_210 / treatment))]	1.000000	0.057143	0.108108
4	[(u_12 / of :1 (u_8 / COORD))]	0.736842	0.050000	0.093645
5	[(u_12 / of :2 (u_21 / condition))]	0.580357	0.232143	0.331633
6	[(u_65 / for :2 (u_210 / treatment))]	1.000000	0.050000	0.095238
7	[(u_8 / COORD :0 (u_57 / improve))]	1.000000	0.039286	0.075601
8	[(u_12 / of :1 (u_117 / component))]	0.916667	0.039286	0.075342
9	[(u_8 / COORD :0 (u_21 / condition) :0 (u_15...	0.769231	0.035714	0.068259

is_treat

u_0 / against :2 (u_1 / condition)

- appears in 18 sentences
- example:

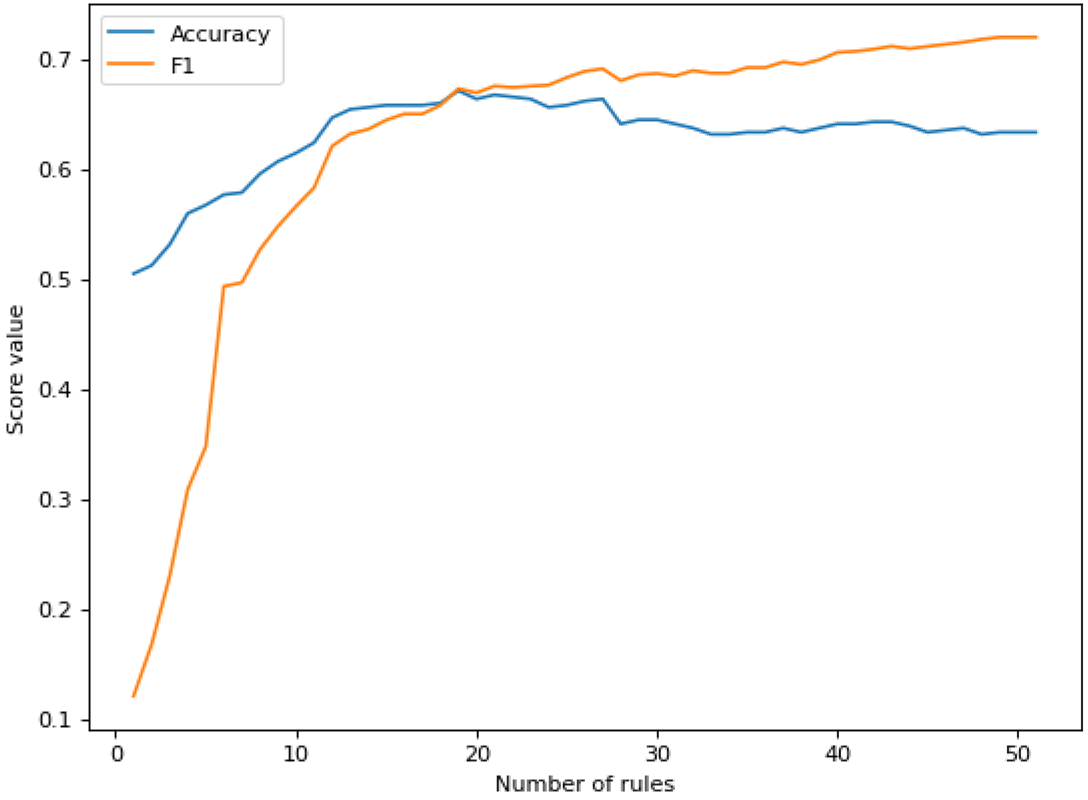
„influence is a promising functional food against condition, owing to a variety of bioactive compounds"



is_treat

Model
- trainer gave 50 features
- ruleset: increment by 1 rule ordered by feature importance
- best results: 28 features

is_treat	POTATO min_edge = 1, 4lang
accuracy	0.5932
precision	0.6154
recall	0.7272
F1	0.6667



is_cause

Trainer 1

- **min_edge = 0**
  -> we can have tokens as rules
- top 10 features:

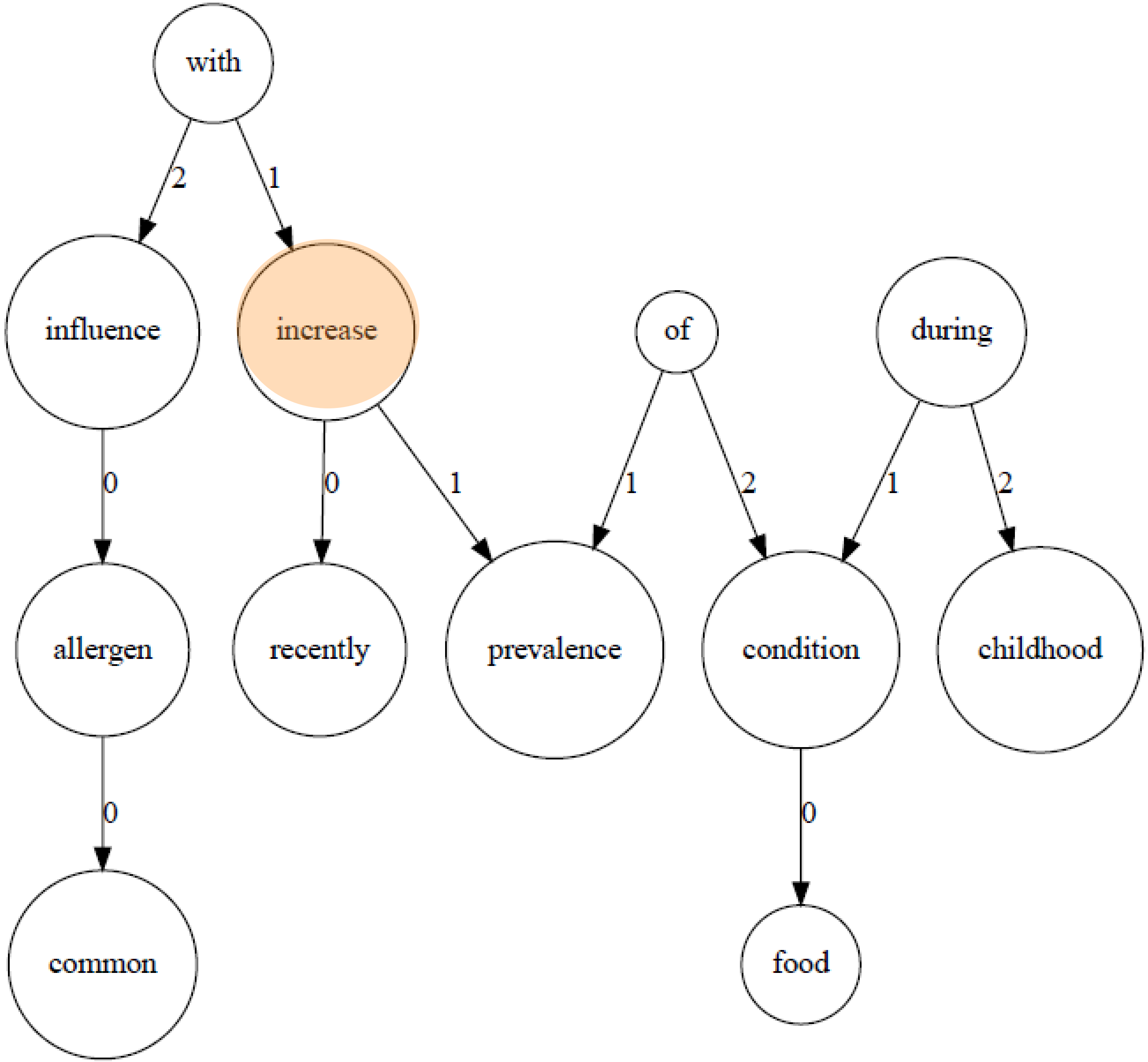
	Feature	Precision	Recall	Fscore
0	[(u_740 / increase)]	0.666667	0.169492	0.270270
1	[(u_19 / patient)]	0.428571	0.076271	0.129496
2	[(u_66 / symptom)]	0.500000	0.050847	0.092308
3	[(u_363 / protein)]	0.437500	0.059322	0.104478
4	[(u_161 / important)]	0.400000	0.050847	0.090226
5	[(u_110 / high)]	0.407407	0.186441	0.255814
6	[(u_8 / COORD :0 (u_106 / product))]	0.461538	0.050847	0.091603
7	[(u_460 / among)]	0.600000	0.076271	0.135338
8	[(u_167 / population)]	0.500000	0.050847	0.092308
9	[(u_168 / child)]	0.538462	0.059322	0.106870

is_cause

increase

- appears in 30 sentences
- example:

„recently, the prevalence of food condition during childhood is increasing, with influence being common allergens"

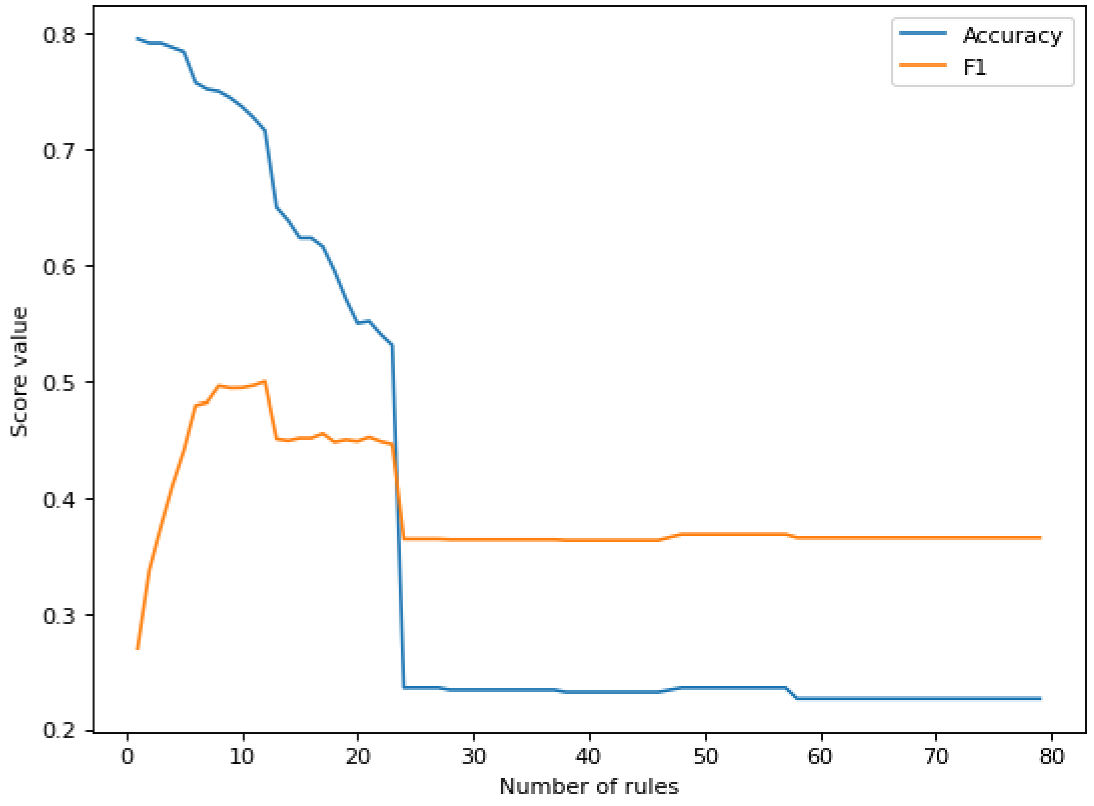


is_cause

Model

- trainer gave 80 features
- ruleset: increment by 1 rule ordered by feature importance
- best results: 8 features

is_cause	POTATO min_edge = 0, 4lang
accuracy	0.5932
precision	0.6154
recall	0.7272
F1	0.6667



is_cause

Trainer 2

- **min_edge = 1**
  -> we don't want tokens
- top 10 features:

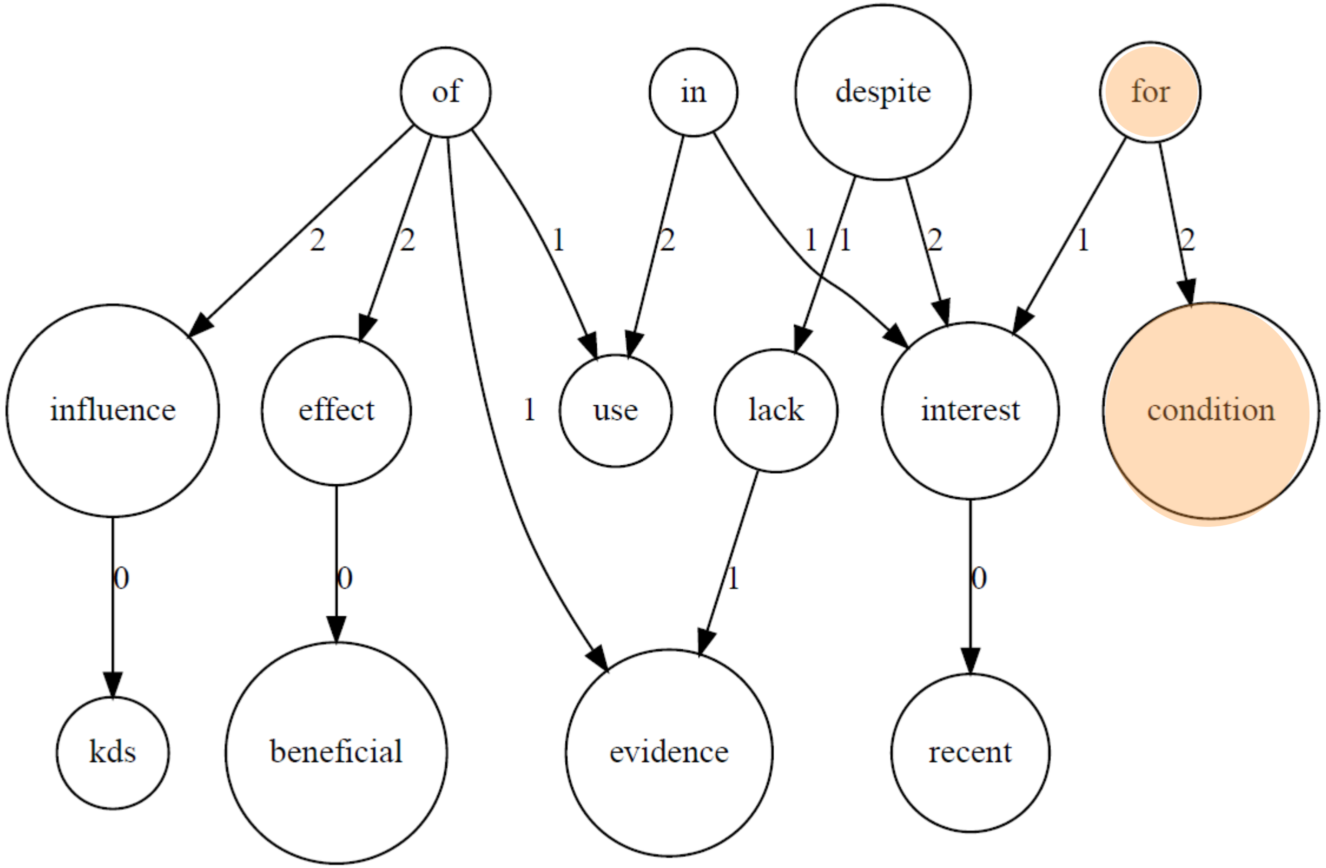
	Feature	Precision	Recall	Fscore
0	[(u_8 / COORD :0 (u_106 / product))]	0.461538	0.050847	0.091603
1	[(u_65 / for :2 (u_21 / condition))]	0.333333	0.059322	0.100719
2	[(u_8 / COORD :1 (u_17 / influence))]	0.352941	0.050847	0.088889
3	[(u_8 / COORD :0 (u_5 / disease) :0 (u_21 / ...	0.304348	0.059322	0.099291
4	[(u_277 / factor :0 (u_153 / risk))]	0.636364	0.059322	0.108527
5	[(u_12 / of :1 (u_179 / development))]	0.333333	0.042373	0.075188
6	[(u_4 / with :1 (u_89 / associate))]	0.303571	0.144068	0.195402
7	[(u_12 / of :1 (u_330 / intake))]	0.450000	0.076271	0.130435
8	[(u_18 / in :1 (u_21 / condition))]	0.222222	0.033898	0.058824
9	[(u_8 / COORD :0 (u_5 / disease))]	0.285714	0.084746	0.130719

is_treat

u_0 / for :2 (u_1 / condition)

- appears in 21 sentences
- example:

„despite recent interest in the
use of influence (kds) for condition,
evidence of beneficial effects is lacking"

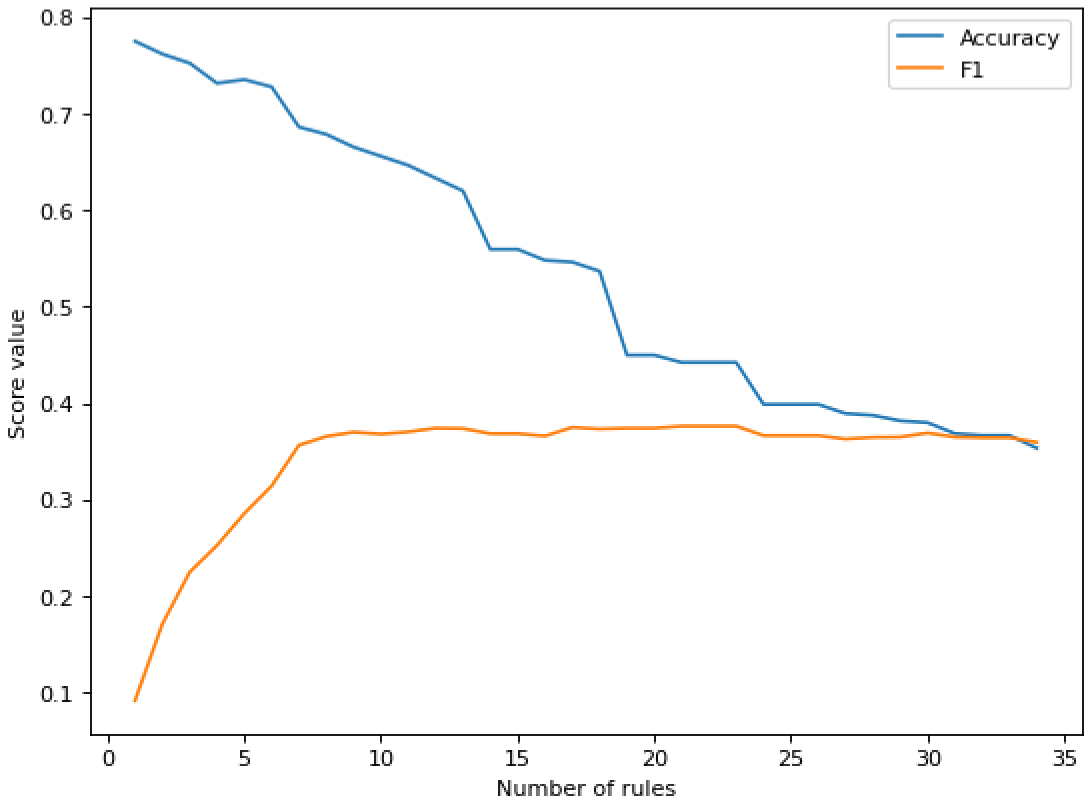


is_cause

Model

- trainer gave 50 features
- ruleset: increment by 1 rule ordered by feature importance
- best results: 9 features

is_cause	POTATO min_edge = 1, 4lang
accuracy	0.5593
precision	0.2272
recall	0.3571
F1	0.2778



Main Findings & Conclusions

White-box model

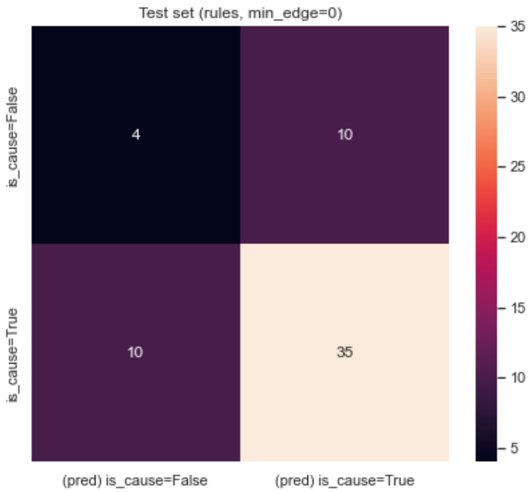
Interpretable and explainable results

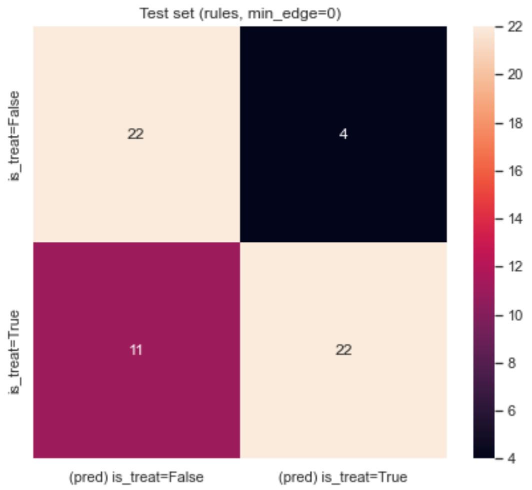
Time

Managing rules and interpreting results is quite time consuming

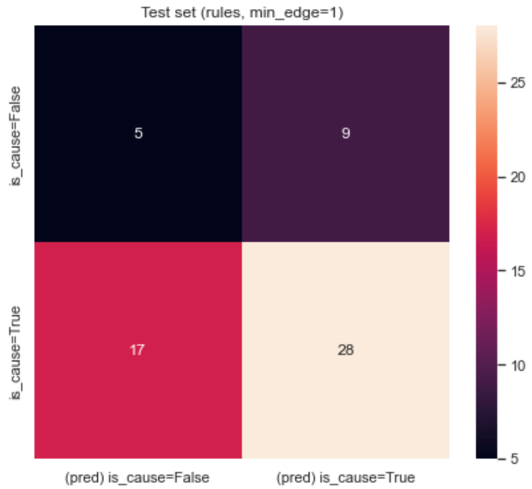
rule based system + ML

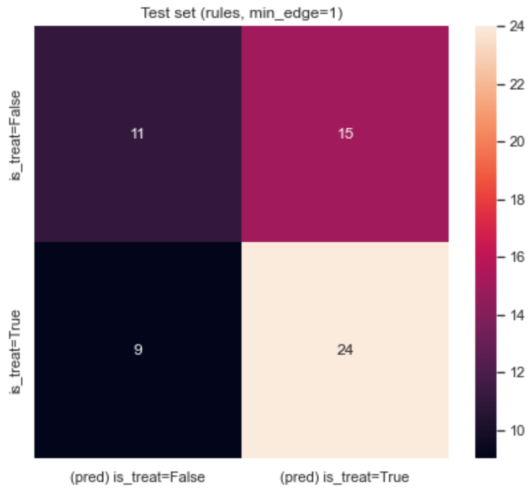
rules w/ tokens





rules w/out tokens





Baseline (full)